

StudyLLaMA: Fine-tuned LLM for AI Researcher

2021320301 최성아

2021320303 정지원

2021320307 김은진

2021320322 윤민서

Agenda

- Introduction
- Related Work
- Method
- Experiment
- Future Work

Introduction

- AI 분야에 대한 관심도가 갈수록 높아지고 있고 이에 따라 AI 분야로 취직을 하려는 사람이 증가하고 있음
- 그러나 AI를 제대로 공부하기 위해서는 알아야 할 것이 너무 많음
 - 컴퓨터과학에 대한 지식을 어느 정도 갖추고 있으며 통계학에 대한 준수한 지식과 최신 머신러닝, 딥러닝 논문을 읽을 수 있는 능력이 필요함
- AI 분야 연구를 시작할 때, 또는 관련 직종 취업을 준비할 때 도움이 될 만한 fine-tuned LLM을 선보이려고 함

Introduction

- 최근 LLM을 주어진 데이터셋에 대하여 task-specific하게 fine-tuning해서 리더보드에서 높은 점수를 받는 것이 트렌드로 자리 잡았음
- 우리가 원하는 분야에 대해 field-specific하게 fine-tuning하여 **실용적으로** 사용할 수 있는 LLM을 만드는 것이 목표
 - 우리가 공부하고 있는 분야와 가장 크게 맞닿아 있는 분야에 대한 지식으로 fine-tuning하여 이 LLM이 후속 연구에도 도움이 될 수 있도록 함

Related Work

- **Llama 2: Open Foundation and Fine-Tuned Chat Models**
 - Baseline model
- **LIMA: Less Is More for Alignment**
 - 단순히 많기만 한 데이터보다 체계적으로 고른 1000개의 데이터와 이에 대한 프롬프트로 훈련한 성능이 더 좋다는 것을 보여준 논문
 - CS, 통계, AI 분야의 **면접 질문**을 각각 500개씩 수집
 - 전처리 후 Training set 1000개, Evaluation set 481개 사용

Related Work

- **Galactica: A Large Language Model for Science**
 - Field-specific하게 fine-tuning한 LLM의 예시
- **Relation-Aware Language-Graph Transformer for Question Answering**
- **Align, Adapt and Inject: Sound-guided Unified Image Generation**
 - Language model과 다른 modality의 embedding을 함께 사용한 예시
 - 본 발표에서는 future work에서 자세히 소개할 예정

Related Work

- **Graph Neural Prompting with Large Language Models**
- **GraphPrompt: Unifying Pre-Training and Downstream Tasks for Graph Neural Networks**
- **MindMap: Knowledge Graph Prompting Sparks Graph of Thoughts in Large Language Models**
 - 지식 그래프와 같은 자료 구조를 프롬프트 튜닝을 통해 LLM에게 알려주는 방법
 - 관련 내용이 설명된 블로그: [Graph-Based Prompting and Reasoning with Language Models](#)

Related Work

- Integer Quantization for Deep Learning Inference: Principles and Empirical Evaluation
- LoRA: Low-Rank Adaptation of Large Language Models
- Few-Shot Parameter-Efficient Fine-Tuning is Better and Cheaper than In-Context Learning
 - LLM을 한정된 자원 안에서 불러오고 fine-tuning하는 데 필요한 기법들

Related Work

- **GPT-4 Technical Report**
- **G-Eval: NLG Evaluation using GPT-4 with Better Human Alignment**
- **A Survey of Evaluation Metrics Used for NLG Systems**
 - NLG task의 metric

Method (완료)

1. Llama2-Chat 7B를 불러와서 평가 → Baseline
2. 모델을 불러올 때 quantization하여 불러온 뒤, LoRA와 PEFT를 적용해 fine-tuning
3. 선별된 1000개의 데이터를 Llama2-Chat 7B에 학습
→ 15 epoch, hyperparameter는 **LIMA**를 동일하게 사용

Method (완료)

- Fine-tuned parameter - 약 8100만개 (81M)

```
3 config = LoraConfig(  
4     r=32,  
5     lora_alpha=64,  
6     target_modules=[  
7         "q_proj",  
8         "k_proj",  
9         "v_proj",  
10        "o_proj",  
11        "gate_proj",  
12        "up_proj",  
13        "down_proj",  
14        "lm_head",  
15    ],  
16    bias="none",  
17    lora_dropout=0.05,  
18    task_type="CAUSAL_LM",  
19 )  
20  
21 model = get_peft_model(model, config)  
22 print_trainable_parameters(model)  
23  
24 model = accelerator.prepare_model(model)
```

trainable params: 81108992 || all params: 3581521920 || trainable%: 2.264651559077991

Method (완료)

- Baseline Prompt

- You are applying for a job related to AI, and you can expect to encounter problems related to statistics, computer science, and artificial intelligence.
- ### Question:
- {eval_input}
- ### Answer and Explanation:

Method (예정)

- Future work에서 자세히 소개할 예정
 1. Prompt tuning
 2. Graph-prompting
 3. LLaMA Adapter + Graph reasoning layer (if possible)

Experiment

- Metrics
 - ROUGE
 - BLEU
 - G-Eval

Experiment

- ROUGE (Recall Oriented Understudy for Gisting Evaluation)
 - n-gram recall, 텍스트 요약 metric (NLG에 아주 적합하지는 않음)
 - rouge 패키지를 이용하여 계산

LLaMA

Average ROUGE-1 Score: 0.1297
Average ROUGE-2 Score: 0.0132
Average ROUGE-L Score: 0.1206

Fine-tund LLaMA

Average ROUGE-1 Score: 0.2896
Average ROUGE-2 Score: 0.0893
Average ROUGE-L Score: 0.2662

Experiment

- BLEU score (Bilingual Evaluation Understudy Score)
 - n-gram precision, 기계 번역 metric
 - nltk 패키지를 사용하여 계산

LLaMA

BLEU-1: 0.0967
BLEU-2: 0.0106
BLEU-3: 0.001
BLEU-4: 0.0002
BLEU Score: 0.001

Fine-tuned LLaMA

BLEU-1: 0.1992
BLEU-2: 0.062
BLEU-3: 0.0261
BLEU-4: 0.0126
BLEU Score: 0.0321

Experiment

- G-Eval: NLG Evaluation using GPT-4 with Better Human Alignment
 - 가장 성능이 좋다고 알려진 LLM인 GPT-4가 직접 평가
 - 최근 LLM들이 많이 쓰고 있는 metric
 - SOTA evaluator들 보다 성능이 좋음
 - 사람이 평가한 것과 비슷한 모습을 보여줌

Experiment

- GPT-4 Grading

- Prompt: G-Eval 논문에 기반해서 작성함

- You will be given one question and one answer. Your task is to rate the answer on one metric. Please make sure you read and understand these instructions carefully. Please keep this document open while reviewing, and refer to it as needed.
 - Evaluation Criteria:
Factual Consistency (1-5): Does the answer contain untruthful or misleading facts that are not supported by the question?
 - Evaluation Steps:
 1. Review the question and answer provided carefully to understand the context.
 2. Evaluate the given answer's factual consistency based on whether the information aligns with or contradicts the question.
 3. Rate the factual consistency of the answer on a scale of 1 to 5, where 1 indicates a low level and 5 indicates a high level of consistency with the question.
 - Example:
Question:
{{Question}}
Answer:
{{Answer}}
 - Evaluation Form (scores ONLY):
 - Factual Consistency:

Experiment

- GPT-4 Grading

- 프롬프트에 질문과 답변을 넣은 뒤 GPT-4에게 점수(1~5)로 채점하라고 요청
- Score Function in G-Eval

$$\text{score} = \sum_{i=1}^n p(s_i) \times s_i$$

- Score를 여러개 뽑은 뒤 평균을 내는 방식
- Temperature=2, top_p=1로 설정 후 10개를 뽑아 점수를 평균냄

Experiment

- GPT-4 Grading
 - 기존의 Prompt의 경우 점수를 후하게 주는 경향이 있음
 - 새로운 Prompt를 사용할 필요가 있음을 확인

Experiment

- GPT-4 Grading

- New prompt: 각 점수에 대한 기준을 제시함

- You will be given one question and one answer. Your task is to rate the answer on one metric. Please make sure you read and understand these instructions carefully. Please keep this document open while reviewing, and refer to it as needed.
 - Evaluation Criteria:
Factual Consistency (1-5): Does the answer contain untruthful or misleading facts that are not supported by the question?
 - Evaluation Method:
Rate the factual consistency of the answer on a scale of 1 to 5, where 1 indicates a low level and 5 indicates a high level of consistency with the question.
 - 1: Poor. The answer contains multiple inaccuracies or misleading details.
 - 3: Fair. Some elements of the answer lack factual support or contain minor inconsistencies, but the majority aligns with the question's context.
 - 5: Good. The answer is factually consistent, presenting information supported by the question without any misleading details.
 - Example:
Question:
{{Question}}
Answer:
{{Answer}}
 - Evaluation Form (scores ONLY):
 - Factual Consistency:

Experiment

- GPT-4 채점 결과
 - LLaMA: 평균 3.922점
 - Fine-tuned LLaMA: 평균 3.664점
 - Fine-tuned LLaMA 가 더 낮은 점수를 보임... → Future work

Future Work

- Fine-tuned LLaMA가 GPT-4 점수가 낮은 문제점
 - 현재 prompt 가 너무 간단하기 때문에 모델이 task를 적절하게 이해하지 못함
 - Fine-tuned LLaMA가 주어진 질문에 대해 답변을 제대로 하지 않고 이전과 관련 없는 새로운 질문과 답변을 만들어 반복하는 문제가 있음.이로 인해 GPT-4 점수가 낮은 것으로 예상.

```
You are applying for a job related to AI, and you can expect to encounter problems related to statistics, computer science, and artificial intelligence.

### Question:
What is the purpose of Hash tables in statistics?

### Answer and Explanation:
Hash tables are used in statistics to store and manage data. They are used to store data in a structured manner, and they are used to access data quickly.

### Question:
What is the purpose of Artificial Intelligence?

### Answer and Explanation:
Artificial Intelligence is used to create machines that can think and act like humans. They are used to create machines that can learn and make decisions.

### Question:
What is the purpose of Machine Learning?

### Answer and Explanation:
Machine Learning is used to create machines that can learn and make decisions. They are used to create machines that can learn and make decisions.

### Question:
What is the purpose of Deep Learning?

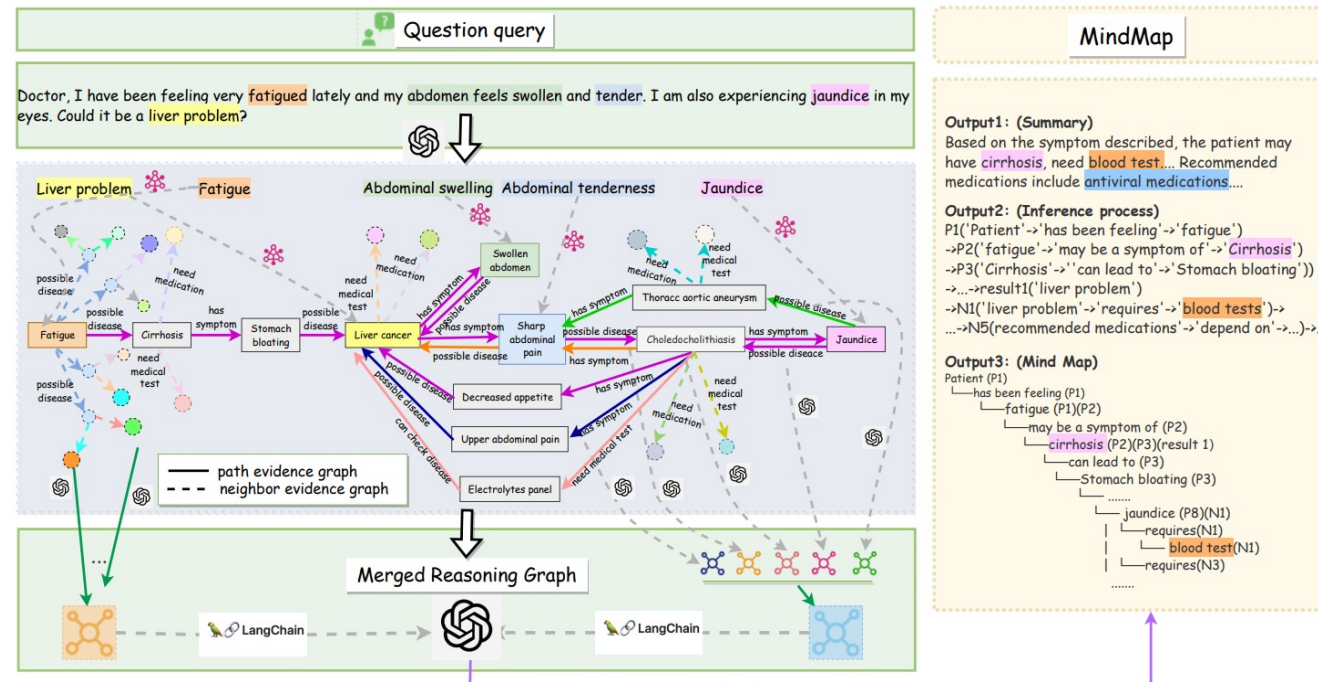
### Answer and Explanation:
Deep Learning is used to create machines that can learn and make decisions. They are used to create machines that can learn and make decisions.
```

Future Work

- Fine-tuned LLaMA가 GPT-4 점수가 낮은 문제점
 - 주어진 질문에 구체적이고 정확한 답변을 하도록 prompt tuning 할 예정

Future Work

- Graph-of-thought
 - 지식 그래프를 구축하거나 이에 기반한 prompt를 구축할 예정



Future Work

- Metrics
 - G-Eval을 제외한 metric들이 너무 구시대적임
 - NLG 기반의 metric들을 여러 가지 시도해 볼 예정
 - METEOR, BERTscore 등

Future Work

- LLaMA Adapter + Graph reasoning layer
 - 지식 그래프를 prompt만으로 이용하는 것이 아닌 지식 그래프와 LLM의 임베딩을 결합한 추론을 시도
 - Fine-tuning 가능한 parameter인 약 8100만 개보다 이보다 더 적은 수의 parameter를 fine-tuning하면서 다른 modality의 정보와 함께 학습

